科研机构的科研状况研究
——基于论文特征项共现分析方法

庞弘燊　方　曙　付鑫金　杨志刚

摘　要　采用论文特征项的共现分析方法，以中科院国家科学图书馆为研究样本，从单个论文特征项、两个特征项的共现到三个特征项的共现进行了分析研究。结果发现，基于论文特征项的共现分析方法可以较好地揭示科研机构的科研状况，如研究的主题领域、研究团体、所发表论文的期刊类型，以及科研人员、研究主题、发文期刊之间的关系等。图6。表3。参考文献5。

关键词　文献计量学　共现分析　多重共现　科研状况

A Method to Reveal the Status of Scientific Research in Institutions Based on the Occurrence of Entities in Papers

Pang Hongshen　Fang Shu　Fu Xinjin　Yang Zhigang

Abstract: This paper uses a method based on the occurrence of entities in papers to analyze the status of research in a sample of National Science Library. The method uses a single entity, two occurrence entities and three occurrence entities for analyzing. The results show that this method can be better to reveal the status of scientific research in institutions, which includes the research topics, research groups, the type of published journals, and the relationship among research staff, research topics and published journals.

Keywords: Bibliometrics; Occurrence and co-occurrence; Multiple occurrence; Research status

1　引言

目前对科研机构的科研评价大都使用一些较宏观的评价指标，如着重评估科研机构的人员职称、发文数量、科研项目等。基于成果式的评价方法使得科研机构过于注重基于成果的科研管理，而忽略了一些微观层面上的科研状况。一般来说，科研机构的科研实力体现在其科研人员、前沿研究主题、发表的高质量论文上，通过分析科研机构的研究主题、研究团体及其发表论文的载体，可以了解到一个科研机构的研究热点、研究团体构成等情况。

本研究拟从微观的层面出发，以中科院国家科学图书馆（以下简称“国科图”，包括设在北京的总馆，以及设在兰州、成都、武汉的三个分馆）[1] 为研究样本，通过不同层面的论文特征项共现分析方法，来揭示其研究主题、研究团体、论文发表期刊及其相互间的联系。科研机构可以根据其分析结果，把握自己的研究主题方向、机构成员的合理配置，或者与其论文发表的期刊建立长期的合作关系等。使用该方法也可对几个类似机构进行对比分析，反映不同机构之间的异同或强弱，有助于促进机构的发展。

2　研究方法与数据来源

2.1　研究方法

对于一个科研机构来说，其发表的论文承载了其大部分的最新科研成果，通过研究科研机构所发表论文的特征项共现分析，可以了解到该科研机构的研究主题、科研人员的配置、以及发表论文的期刊杂志等的情况。因此本研究拟采用如图1所示的一系列方法，揭示出科研机构的研究主题、研究团体、论文发表期刊及其相互联系。

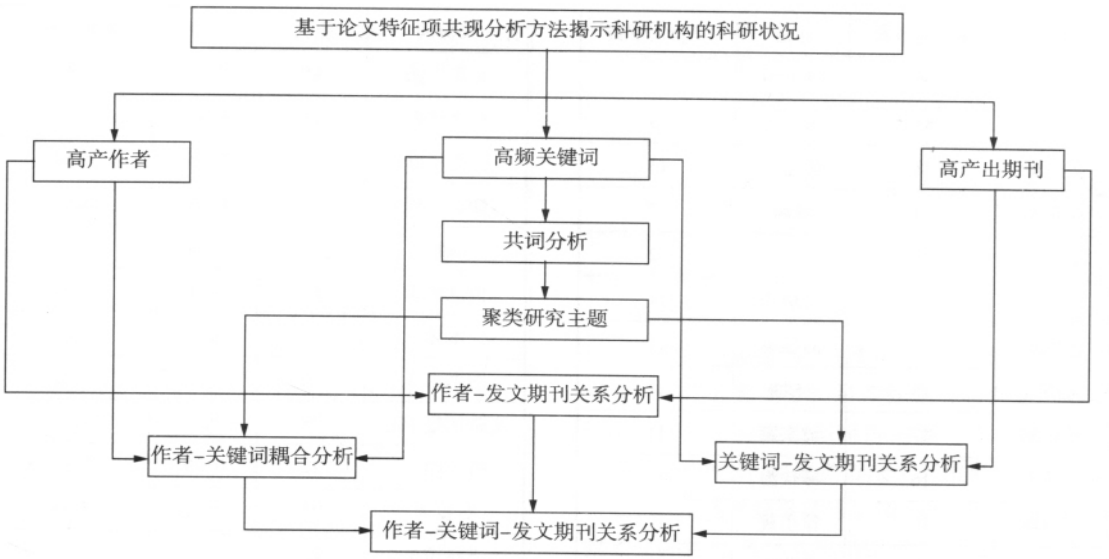

图 1 基于论文特征项共现分析方法揭示科研机构的科研状况

首先通过对科研机构所发表论文的高产作者、高频关键词、高产出期刊进行统计,找出该机构当中发表论文较多的科研人员、研究主要集中的关键词领域、论文主要发表的期刊类型。接着对高频关键词进行共词分析,并聚类出研究主题,找出科研机构的主要研究主题。之后通过作者 - 关键词共现分析,找出科研机构中科研人员的主题研究方向,或找出科研机构中属于同类研究主题的研究团体。同时也通过作者 - 发表期刊分析方法,找出科研人员所偏好发表论文的期刊类型分布。而通过关键词 - 发文期刊的关系分析,可以找出科研机构在期刊上发表某类研究主题的集中程度。最后通过分析作者 - 关键词 - 发文期刊的关系,可以找出具体的科研人员在某类期刊上所集中发表某类研究主题的论文。

2.2 数据来源与处理

在论文数据的搜集上,为便于统计的需要,本研究只选取中文数据库。所检索出来的数据均来自中国知网(CNKI)的中国学术期刊网络出版总库数据库,论文发表年份限定为 2000 - 2009 年,共检索出关于国科图(包括各分馆)的 2493 条记录(检索日期: 2010 年 9 月 23 日)。

然后对检索出来的论文记录进行筛选,排除论文第一单位不是国家科学图书馆的记录以及不属于学术论文的记录等,剩下 1958 条记录。该 1958 条记录可以看作国科图在 2000 - 2009 年间所发表的学术论文数量,在后续的研究当中,将基于这 1958 条论文数据进行分析。

3 数据统计

统计科研机构所发表论文中的高产作者、高频关键词、高产出期刊,可以发现科研机构中的主要研究人员,主要研究主题以及研究成果的主要载体。由于第一作者一般是论文中的主要贡献者,所以本研究统计中所涉及到的作者均指发表论文的第一作者。

通过统计国科图的高产作者(如表 1 所示)、高频关键词(如表 2 所示)和高产出期刊(如表 3 所示),可以看出国科图发文量居于前十的高产作者包括有白国应、文榕生、张晓林、初景利、李春旺、张志强、孟广均、张智雄、吴新年、马建霞。还有可把其它年均发文量达 1 篇以上的研究人员视为国科图的高产作者。

表 1 2000 – 2009 年国科图高产作者表

作者	发表文章数	作者	发表文章数
白国应	96	祝忠明	13
文榕生	85	郭家义	13
张晓林	38	曲建升	13
初景利	33	林曦	12
李春旺	21	孙坦	12
张志强	20	冯瑞华	11
孟广均	18	向桂林	11
张智雄	18	张娴	11
吴新年	17	谭宗颖	11
马建霞	16	冷伏海	11
吴振新	16	徐引篪	11
毛军	15	吕俊生	10
金碧辉	15	李景	10
黄国彬	13	孙成权	10

在高频关键词的统计上,选取出现频次 20 或以上的关键词为高频关键词,其累计百分比达到了 10.6%,表 2 中的高频关键词可视为国科图的研究重点。其中,数字图书馆居于首位,可见国科图一直致力于数字图书馆理论与实践的研究,成果较多。同时国科图在信息服务、文献分类、知识管理、元数据等的研究上也具有较多的成果,其相关的关键词频次都居于前十位。除此以外,国科图在信息检索、本体、学科馆员、开放获取、纳米技术、情报研究、长期保存、科技期刊、文献计量学等方面的研究上,也有一定的成果,其相关的关键词频次也居于前列。

表 2 2000 – 2009 年国科图发表论文的高频关键词表

续表

关键词	频次	百分比	累积百分比
数字图书馆	135	1.44	1.44
图书馆	95	1.01	2.45
信息服务	55	0.59	3.04
文献分类	53	0.56	3.6
中国科学院	51	0.54	4.14
分类方法	45	0.48	4.62
分类体系	45	0.48	5.1
分类标准	41	0.44	5.54
知识管理	37	0.39	5.93
元数据	35	0.37	6.3
信息检索	34	0.36	6.66
本体	32	0.34	7
图书馆学	30	0.32	7.32
文献情报中心	28	0.3	7.62
被引频次	27	0.29	7.91
数据库	25	0.27	8.18
学科馆员	25	0.27	8.45
开放获取	25	0.27	8.72
纳米技术	24	0.26	8.98
情报研究	24	0.26	9.24
长期保存	23	0.25	9.49
科技期刊	22	0.23	9.72
可持续发展	22	0.23	9.95
文献计量学	21	0.22	10.17
图书馆员	21	0.22	10.39
知识服务	20	0.21	10.6

在论文发表期刊上,选取发文量达 20 篇或以上的发文期刊为高产出期刊(如表 3 所示),相当于国科图每年在该类期刊上至少发表了 2 篇论文。国科图大量的研究成果发表于其下属出版中心所发行的杂志《图书情报工作》和《现代图书情报技术》上。除此以外,国科图发表的论文也较多见于图书馆学、情报学领域以及相关科学领域的核心期刊上,如《图书馆杂志》、《图书馆理论与实践》、《图书馆建设》、《情报杂志》、《科学观察》、《情报科学》、《情报理论与实践》、《图书馆论坛》、《中国图书馆学报》、《大学图书馆学报》、《地球科学进展》、《新材料产业》等。

表 3 2000 – 2009 年国科图高产出期刊表

期刊名	发文量	期刊名	发文量
图书情报工作	380	中国图书馆学报	44
现代图书情报技术	200	大学图书馆学报	41
现代情报	98	地球科学进展	38
图书馆杂志	76	新材料产业	38
图书馆理论与实践	65	图书与情报	37
图书馆建设	65	图书情报知识	27
情报杂志	60	图书馆工作与研究	27
科学观察	54	江西图书馆学刊	22
情报科学	53	情报学报	22
情报理论与实践	52	中外科技信息	22
图书馆论坛	50	世界科技研究与发展	20

4 共词分析

共词分析方法属于内容分析方法的一种。它的原理主要是对一组词两两统计它们在同一篇文献中出现的次数,以此为基础对这些词进行聚类分析,反映出这些词之间的亲疏关系,进而分析这些词所代表的学科和主题的结构变化。它利用大量文献中共同出现的关键词对有效地反映文本关键词之间的关联强度,减少了关键词的空间,用一套结构图有效地展示了关键词之间的关联[2]。共词分析方法可以发现科研机构的研究热点及其研究主题的结构。

4.1 抽取高频关键词

在所检索到的 1,958 篇论文中,出现的关键词共有 5,107 个,截取出现频次在 6 次(含)以上的 219 个关键词,其累积频次占总频次的 30.48%,将这 219 个关键词作为国科图的研究热点。这些关键词在国科图所发表的论文中出现的频率较高,在一定程度上体现了国科图的研究热点。

4.2 形成共词矩阵和相似(相异)矩阵

仅按出现的频次对这些关键词进行线性排序,还不足以全面反映它们之间的关系。为此,根据共词的原理,对这些关键词做进一步处理,两两统计它们在同一篇论文中出现的次数。如果两个关键词同时出现的频率高,说明它们之间的关系密切,采用自编的词频统计软件对共词进行统计,形成了一个 219 × 219 的共词矩阵,对角线上取值为 0。

本研究中选取余弦相似性测度(Cosine similarity measure)的方法形成相似矩阵,计算共词向量间的余弦,其公式为[3]: $Cosine(x,y) = x_i y_i / \sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}$ 在共词矩阵的基础上,通过 SPSS 软件形成余弦系数相似矩阵。

4.3 聚类分析

聚类分析是一种重要的多元统计方法,也是文献计量研究中经常用到的数据结构挖掘技术,通过对共词矩阵的聚类分析可以挖掘学科知识的微观结构[4]。本研究在相似矩阵的基础上,通过 SPSS 形成树形聚类图,在聚类方法上采用聚类分析中应用最为广泛的系统聚类法(Hierachical cluster)。在聚类分析上,选择欧几里得距离平方和作为变量距离的测度方法,类间距离的计算方法采用组间联接法。通过聚类把高频关键词分成 11 个大类: 数字图书馆建设、信息服务研究、文献分类法研究、国科图相关机构及图情事业的情报研究、信息系统研究、科技期刊研究、科学计量学研究、信息资源建设与情报服务、科学技术发展态势研究、未定义 1(该类中包含有各类主题的关键词,故不作定义)、用户服务研究。

5 共现分析

5.1 作者 – 关键词耦合分析

作者 – 关键词耦合分析方法,是指利用作者文献集关键词的耦合强度分析作者之间关系的一种方法[5]。笔者认为通过作者 – 关键词耦合分析法可以找出科研机构中的主题研究方向及其相应的研究团体。

首先分析高产作者与高产关键词之间的关系,可以发现国科图主要研究人员的主要研究领域,如果截取高产作者 – 高频关键词矩阵太大,会形成稠密的 2 模图,图形可视化的效果不大好,故选取 2000 – 2009 年间国科图前 25 的作者以及前 107 个高频关键词,形成高产作者 – 高频关键词关系矩阵(25 × 107),通过 2 模网络对其进行可视化,从中可以看出高产作者的具体热门研究领域。

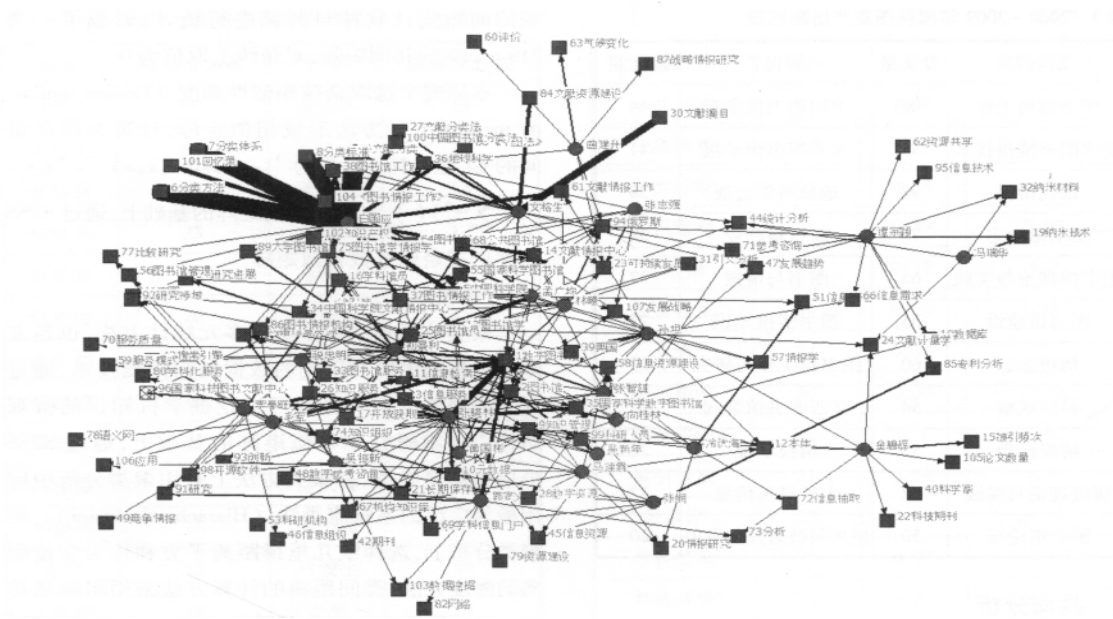

图 2 高产作者 – 高频关键词 2 模网络图

在高产作者 – 高频关键词 2 模网络图中能够看出这些高产作者的具体研究领域,越靠外围的作者研究领域较专,而越靠中心的作者研究领域越广,居于中心的关键词是多个作者所共同研究的主题领域。图中连线的粗细,代表作者采用该关键词的频次,作者采用关键词频次越高,作者和关键词之间的连线则越粗。从图 2 中可以看出大部分高产作者的具体研究领域有一定的重合之处,但有一些处于图中外围位置的作者,其研究的主题比较独特,和其它作者的论文关键词重合较少,如白国应和文榕生在文献编目和分类上有专门的研究,曲建升在地球科学战略情报研究上较多,谭宗颖和冯瑞华的研究较多集中在纳米技术的分析上,金碧辉的研究领域主要集中在计量学上。而其它作者的研究范围都比较广泛,较难从图中识别出来作者的具体研究领域。从居于中心的数字图书馆关键词来看,国科图有多位高产作者从事该方面的研究,而其中以张晓林最为突出。

对作者 – 研究主题的关系进行分析,可以发现在主题研究上所对应的研究团体,有利于科研机构中研究人员的合理配置。笔者选取 2000 – 2009 年间的国科图的发文作者共 159 人(发文量在 4 篇或以上)和高频关键词 219 个,并通过自编软件统计形成作者 – 关键词矩阵(159 × 219)。然后在合并 11 类研究主题的关键词后,形成作者 – 研究主题的关系矩阵(159 × 11),并将其转化成 2 模网络进行可视化,如图 3 所示。图中连线代表作者与该研究领域的关系,连线的粗细代表作者论文中含有该研究主题的关键词频次,如果作者所发表论文包括较多该领域的关键词,那么连线就越粗,表明作者涉及到该主题领域的研究越多。从图 3 可以看出,国科图在数字图书馆建设、信息系统研究、信息资源建设与情报服务、信息服务研究、国科图相关机构及图情事业的情报研究领域中,研究人员较多,并且有部分研究人员的研究跨度多个相关的领域,而在科学技术发展态势研究、科学计量学研究的领域当中,研究人员较少。除此以外,通过图中还能发现研究领域当中的一些主要研究人员,如白国应有大量关于文献分类法的研究;而张晓林在数字图书馆建设领域有较多的研究;在科学计量学的研究方面,金碧辉和张克菊研究成果也较多;在科学技术发展态势的研究上,朱相丽也发表了较多的成果。

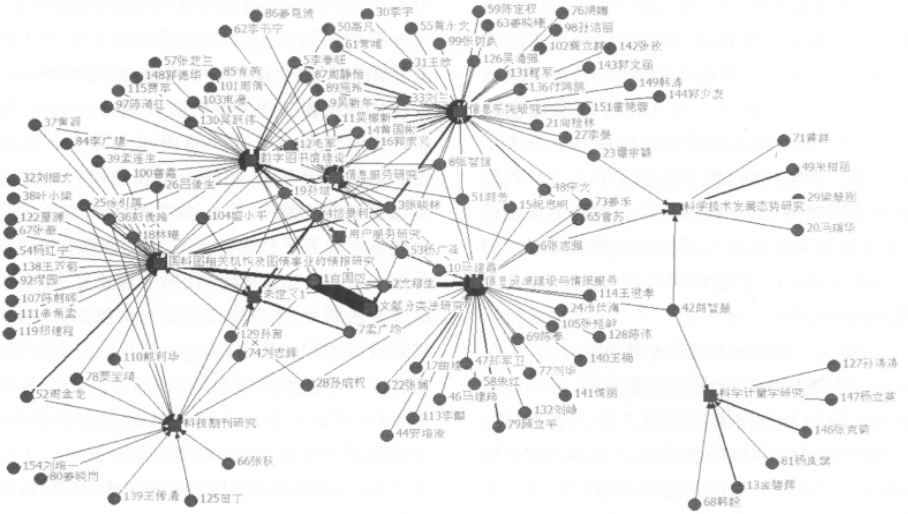

图3 作者 – 研究主题2模网络图

5.2 作者 – 发表期刊关系分析

通过作者 – 发表期刊的关系分析,可以发现科研机构的研究人员在哪类期刊上发表论文较多,或者可以发现科研机构在某类期刊上的稳定作者群体。笔者选取2000 – 2009年间的国科图的发文作者共159人(发文量在4篇或以上),发表期刊51种(发文量在5篇或以上),通过自编软件统计形成作者 – 发表期刊矩阵(159 × 51),并将矩阵转化成2模网络进行可视化,如图4所示。图中连线代表作者与发表期刊的关系,连线的粗细代表作者在期刊上的发文量,如果作者在期刊上所发表论文越多,那么作者与该期刊的连线就越粗。从图4可以

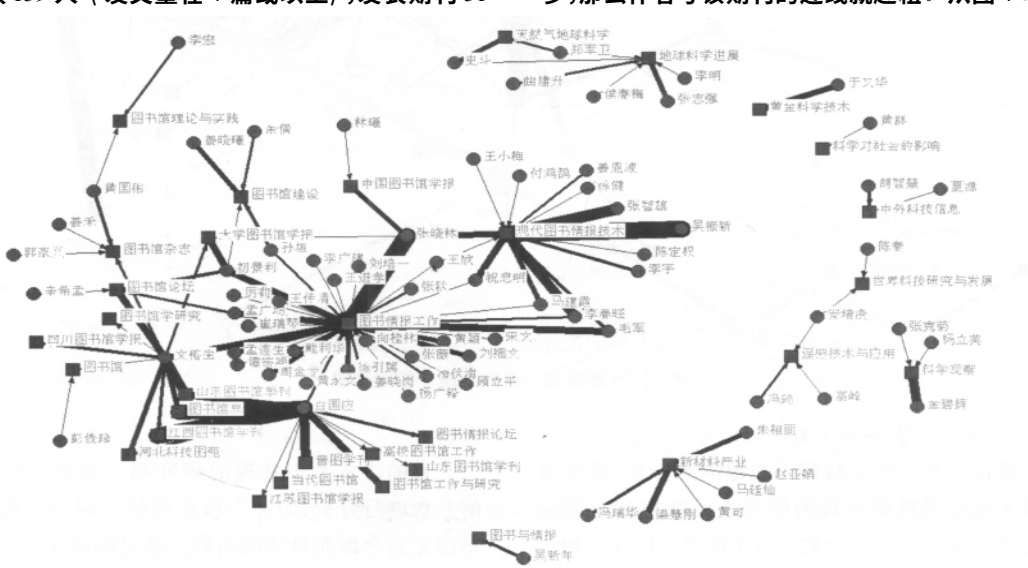

图4 作者 – 发表期刊2模网络图

看出在《图书情报工作》和《现代图书情报技术》国科图具有大量稳定的作者群,国科图的研究人员在这两本刊物上发表了众多的研究成果,其次,在《新材料产业》上也发文较多。而从作者的层面上来看,白国应和文榕生则在众多的图情期刊上发表了大量的文章。张晓林在《图书情报工作》、《现代图书情报技术》和《中国图书馆学报》上发表了众多的论文,张智雄和吴振新的大部分论文都发表在《现代图书情报技术》上,金碧辉则在《科学观察》上发表了较多的论文等。

5.3 发表期刊 – 关键词的关系分析

通过发表期刊 – 关键词的关系分析,可以发现科研机构在期刊上所发表的研究主题。笔者选取 2000 – 2009 年间的国科图发表论文的期刊 51 种(发文量在 5 篇或以上),高频关键词 219 个,通过自编软件统计形成发表期刊 – 关键词矩阵(51 × 219),然后在合并 11 类研究主题的关键词后,形成发表期刊 – 研究主题的关系矩阵(51 × 11),并将其转化成 2 模网络进行可视化,如图 5 所示。图中连线代表发表期刊与该研究领域的关系,连线的粗细代表国科图在所发表的期刊论文中含有该研究主题的关键词频次,如果所发表的期刊论文包括较多该领域的关键词,那么连线就越粗,表明国科图在该期刊上发表了较多涉及该主题领域的研究论文。从图 5 可以看出,在国科图所发表论文中,有关于文献分类法研究、信息资源建设与情报服务、数字图书馆建设等主题的论文都发表在多种的图情期刊上;而关于科学技术发展态势的研究方面,大量论文可见于《新材料产业》;在科学计量学方面的研究,更多的成果是发表在《科学观察》杂志上。从期刊的角度来看,国科图在《图书情报工作》上发表了多个主题研究领域的论文。

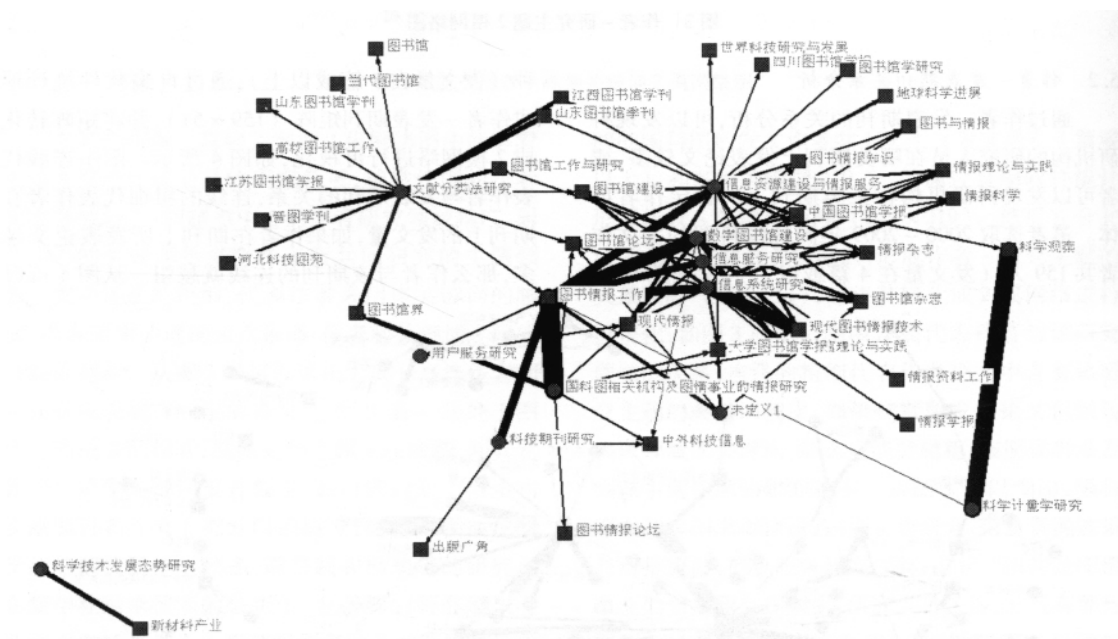

图 5 发表期刊 – 研究主题 2 模网络图

5.4 作者 – 发表期刊 – 关键词分析

通过作者 – 发表期刊 – 关键词的分析,能够更为深入地发现科研机构的研究人员在期刊上所发表的某类研究主题的论文。由于该项分析涉及到三个特征项的共现,因此要通过构建三维矩阵进行分析。

目前许多关于共现的研究都只对两个特征项的共现进行分析,而多个特征项的共现分析能够揭示出更为全面而具体的内容。通过构建作者 – 发表期刊 – 关键词之间关系的三维矩阵,可以形象地揭

示哪类作者在哪类期刊上发表哪类主题的论文。由于多维矩阵涉及到数据量统计庞大和计算复杂的缘故,因此笔者只选取其中的小样本进行分析,作者包括张晓林、初景利、李春旺、张志强四位作者,发表期刊选取了国科图的 51 种高产期刊,关键词则选择 219 个高频关键词,然后通过矩阵转化成作者 - 发表期刊 - 关键词的关系网络图(如图 6 所示)进行分析。图中作者与期刊的连线代表作者在该期刊上的发文量,发文量越多,作者和期刊间的连线则越粗; 而期刊与关键词之间的连线代表作者在该期刊上发表论文时所采用关键词的频次,如果作者在期刊上发表的论文含有较多频次的关键词,则期刊与该关键词之间的连线越粗。

从图中可以看出张晓林主要在《图书情报工作》、《现代图书情报技术》、《中国图书馆学报》三个期刊上发表关于数字图书馆研究的论文,而初景利则主要在《图书情报工作》、《大学图书馆学报》、《图书馆论坛》、《图书馆建设》等期刊上发表论文,并且在《图书情报工作》发表了较多关于图书馆服务研究的论文。李春旺在《现代图书情报技术》发表了较多以数字图书馆技术为研究主题的文章。张志强则主要在《地球科学进展》发表了以地球科学为主题的文章。

笔者把这种出现两个以上特征项共现的情况称作"多重共现"情况,通过分析多重共现现象可以揭示出更为深入的信息内容,比如在分析作者 - 发表期刊 - 关键词这三个特征项共现的关系中,如果从作者的视点出发,可以发现张晓林偏好在《图书情报工作》期刊上发表数字图书馆研究主题的论文,初景利偏好在《图书情报工作》期刊上发表图书馆服务主题类的论文; 从期刊的视角出发,可以发现某类期刊上的稳定作者群,还有其在期刊上所发表论文的主题方向; 而从关键词的角度出发,可以发现发表关于某主题类论文的作者群体和期刊集合。可见多重共现现象比一般的共现现象更为复杂,并且能揭示出更为详细的信息内容,但关于多重共现所能揭示的具体信息内容还有待深入的研究。

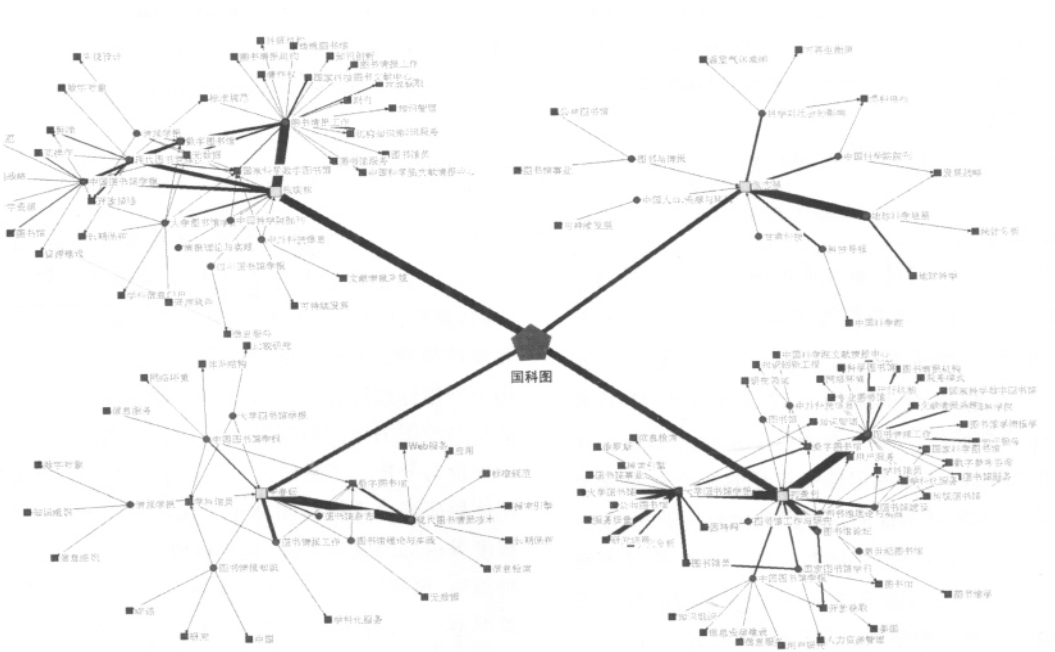

图 6 作者 - 发表期刊 - 关键词的关系网络图

(下转第 88 页)

大量出现在了博士论文的表述中。虽然笔者在分类中也使用了类似概念,但也只是为了方便归类,并非认同每种方法都是博士学位论文中应该表述的。笔者认为博士学位论文中应当阐述的方法是针对论文的主要研究内容所采取的具体、关键方法和技术,这样才能让读者明白论文的研究内容是围绕着什么展开的。

从以上两个问题,我们可以发现二者似乎存在矛盾之处:研究方法的表述是层次分明、全面,还是具体、关键?笔者认为,博士学位论文研究方法的阐述应该是层次分明的,同时突出关键方法与技术。一篇良好的研究方法表述,应当让人明白作者是基于什么思想、采用什么研究方法和具体技术进行研究的。

参考文献

1 P. Dunleavy. 博士论文写作技巧 [M]. 大连:东北财经大学出版社,2009:4.

2 风笑天. 社会学研究方法 [M]. 北京:中国人民大学出版社,2009:159.

3 张寒生. 当代图书情报学方法论研究 [M]. 合肥:合肥工业大学出版社,2006:36.

4 小木虫. 系统分析的概念 [OL]. [2010-06-18]. http://xmujpkc.xmu.edu.cn/zckx/doc/12.ppt.

5 张晓林. 信息管理学研究方法 [M]. 成都:四川大学出版社,1995:103.

6 张寒生. 当代图书情报学方法论研究 [M]. 合肥:合肥工业大学出版社,2006:31.

7 邱均平,王曰芬. 文献计量内容分析法 [M]. 北京:国家图书馆出版社,2008:4.

8 维基百科. 实验 [OL]. [2010-06-18]. http://zh.wikipedia.org/zh/%E5%AE%9E%E9%AA%8C.

9 邱均平,王曰芬. 文献计量内容分析法 [M]. 北京:国家图书馆出版社,2008:1.

10 邱均平等. 信息计量学 [M]. 武汉:武汉大学出版社,2007:26.

11 张晓林. 信息管理学研究方法 [M]. 成都:四川大学出版社,1995:117.

12 张晓林. 信息管理学研究方法 [M]. 成都:四川大学出版社,1995:127.

13 邱均平,王曰芬. 文献计量内容分析法 [M]. 北京:国家图书馆出版社,2008:12.

14 张寒生. 当代图书情报学方法论研究 [M]. 合肥:合肥工业大学出版社,2006:128.

15 风笑天. 社会学研究方法 [M]. 北京:中国人民大学出版社,2009:8.

16 张晓林. 博士论文研究的选题、开题中需要注意的问题 [OL]. [2010-06-18]. http://www.las.cas.cn/jypx/yjsjy/pyyxw/lwyjhj/200909/t20090908_2468576.html.

(杨志刚 中国科学院国家科学图书馆 2009 级图书馆学博士生)

收稿日期:2010-09-10

(上接第 73 页)

参考文献

1 馆情介绍 [EB/OL]. [2010-9-28]. http://www.las.cas.cn/gkjj/.

2 冯璐,冷伏海. 共词分析方法理论进展 [J]. 中国图书馆学报,2006(2):88-92.

3 周静怡,孙坦,陈涛. 共词可视化:以人类基因组领域为例 [J]. 情报学报,2007(4):532-537.

4 杨立英. 科技论文共现理论与应用 [D]. 北京:中国科学院文献情报中心,2007:55.

5 刘志辉. 作者关键词耦合分析及其在研究领域分析中的应用研究 [D]. 北京:中国科学院文献情报中心,2010:2.

(庞弘燊 中国科学院国家科学图书馆 2009 级情报学博士生,方 曙 中国科学院国家科学图书馆成都分馆馆长,付鑫金 中国科学院国家科学图书馆 2009 级情报学博士生,杨志刚 中国科学院国家科学图书馆 2009 级图书馆学博士生)

收稿日期:2011-01-15